Data Visualization - Group 11

Tobias Vittrup Bak - tobib21@student.sdu.dk
Thor Malmby Jørgin - tjoer21@student.sdu.dk
Kevin Torp - ketor21@student.sdu.dk
Philip Schwartz - pschw21@student.sdu.dk

Links to dashboard (Both the links goes to the same dashboard):

- http://www.datavis.cryptobot.dk
- http://207.127.89.16/

Abstract

This data visualization project explores the major safety events, which happened in the public infrastructure of the United States between 2014 and 2022. The dataset consists of 82916 records and have been visualized using Plotly. The dashboard can be seen here and aims at answering the following questions:

1. Is there an increase or decrease in certain types of accidents in the last 9 years?
2. Is there a relation between time periods in the day and certain types of accidents?
3. Do certain types of accidents occur more often in certain environments?

Visualizations on the dashboard reveal tendencies such as a rise in most types of safety events with a large increase in assaults, a connection between business hours and events happening and some states experiencing a higher percentage of specific event types.

In conclusion, the dashboard provided by the project is a valuable tool for understanding and interpreting major safety events in the United States' public transport system.

Table of Contents

Abstract **0**
Table of Contents **1**
1. Background and Motivation **2**
2. Project Objectives **2**
3. Data **3**
3.1 Data processing: 3
4. Visualization/Dashboard **4**
4.1 Design: 4
4.2 Graphs 5
5. Story/Results **11**
5.1 The Story 11
5.2 What surprised us? 12
5.3 What did we learn about the data? 12
5.4 How well did the visualization work, and how could it be further improved? 12
6. Conclusion/Discussion **13**
7. Participation **14**
References: **15**

1. Background and Motivation

The dataset seemed relevant and interesting since it really impacted the people in the dataset, because it is related to accidents in the USA. The dataset also has quite a lot of observations and data to be analyzed.
The dataset's interest lies in its direct connection to people's lives. Exploring and visualizing the data became not just an academic exercise but a chase to understand and communicate the broader link of accidents in the USA.

We were motivated to choose this dataset and visualize it because we find it very interesting to research and find out how to visualize data correctly for people to gain an understanding of the underlying data.

Visualizing complex datasets is a definite skill that is very important and valuable for companies or governments for decision making. For example, when they are searching for better options or cost cutting opportunities.

2. Project Objectives

The aim of this data visualization project is to gain valuable insights into public traffic accidents in the USA over the last 9 years. The dataset selected for analysis provides a comprehensive view of accidents, with different factors such as time, location and other factors. The dashboard aims at answering the following questions through the visualization:

1. Is there an increase or decrease in certain types of accidents in the last 9 years?
    a. What is the distribution of accident types in this period?
    b. How has it changed over the last 9 years?
    c. How does this differ between states?
2. Is there a relation between time periods in the day and certain types of accidents?
    a. When in the day do people get hurt?
    b. Do we see the same trends in fatal accidents?
    c. How does this change over the course of a year?
3. Do certain types of accidents occur more often in certain environments?
    a. Where on the map do the most accidents happen?
    b. In which states does each event type make up the largest portion of events in the state?
    c. What states have reported the most accidents, per capita?

Our primary goal is to effectively visualize this large-scale dataset in a way that effectively conveys the information to users. Informing them, enabling them to make decisions or form conclusions from the visualization. We aim at making the dashboard useful whether it's a government agency or random individuals stumbling upon the dashboard.

We have kept large attention on user design that aims to streamline the process of information gathering. We vision a user-friendly interface where people, regardless of their level of data knowledge and understanding, can effortlessly navigate through the visualizations to extract meaningful insights about the data.

We believe that conveying the information works better if we not only show the data to people with images, but provide them with an interactive playground that can be zoomed and filtered to the needs of the person. We aim to put the interactive plots in the process of understanding, such that the user can interact with the plots while reading.

3. Data

The data was collected by the U.S. Government. We downloaded the data through their portal at: https://catalog.data.gov/dataset/major-safety-events

The dataset contains entries for major safety events that have happened in the United States public transport system, in the period 2014-2023. There are 82916 records.
For our visualizations the important variables include:

- **Event Date and Event Time:** These variables record the day and time, which is the exact time the accident/major safety events occurred.
- **Event Type Group:** This is used to indicate the type of event that is recorded. This is a categorical variable whose values can be seen in the dashboard. The most common type is Collision.
- **Total Injuries:** This column indicates how many people were injured in the accident. There are also columns for minor and major injuries. This column is the sum of those two. The variable is discrete numerical.
- **Total Fatalities:** This column indicates how many people died in the accident. The variable is discrete numerical.
- **Latitude** and **Longitude:** These columns together indicate the location of the accident. These variables individually are continuous numerical, but they are not suited for plotting in a scatter plot, due to their meaning.

3.1 Data processing:

We removed some columns, which were completely empty, such as hazardous. We only removed columns that we would not be using for the visualizations or the parsing either. The dataset was somewhat messy, since it consists of aggregated data from several counties. Therefore there were more than 100 columns, some with small details, that only some counties used, and only for some types of safety events. Additionally we have removed all columns which were submitted in 2023, as 2023 is still ongoing and we did not want to make any assumptions and skew the data.

We have modified some of the rows. The rows that we modified had the event-type “Non-RGX collision” which was only used by one county, in a couple of rows. These rows showed up as obvious outliers in the dataset, and since a Non-RGX collision is a Collision, the event type of these accidents were changed to Collision.

We created a script for cleaning the data, which saves it to a new, smaller and cleaner, version of the .csv file.

For normalizing the data per capita, each event in our dataset has to be mapped to a location where we know the number of citizens. The only parameters to do so are longitude, latitude and NTD ID. At first the longitude and latitude were used to map each event to a state and county. However, tools for this mapping are few and the only good one was through an api, which limits requests to 1 per second. Our dataset is 80.000+ rows so that would take too long[1]. Therefore, we used a dataset containing all NTD ID's and the state, which they are in.[2] This was coupled with a third dataset, that provided data from the US census, about the population of the states throughout the years. This is used to add a new column that contains a fraction which is this event scaled by the number of citizens in the state. The dataset used can be found here.[3]

Based on the columns within the dataset, additional columns were added. This was done to increase the performance of the application by lessening the amount of calculation needed to show the graphs. From the event date and event time columns the following columns were created: "Hour" and "Month". These columns could then be used to see how many accidents happened at a specific month or hour. Additional columns created from the dataset were: "Event per million citizens"[4], "State" and "Percentage of total accidents in state". The "Event per million citizens" allowed us to sum this column for the events within an area and see how many events happened per million citizens within that state of each type. The "State" column is the state in which the safety event occurred. The "Percentage of total accidents in state" is the percentage that a given event type makes up in that state. This column is also summed.

4. Visualization/Dashboard

4.1 Design:

The dashboard is designed to be similar to how articles are laid out. This means that each page is set up with text and graphs being in the middle 60% of the screen and the pages are scrollable. To the left and right of the graphs, options such as radio buttons, checklists and dropdown menus are placed, which should make them easily accessible to the users. These options are used to modify the elements visualized by the graphs. We use text to explain the context of the graph, and guide the user to what the graphs show. i.e. we strive to help the user in understanding the graph, to prompt their knowledge, and help them while looking at the graph.

The color scheme used is color blind friendly[5] and consists of the following colors: #FFAB66, #1A6EB3, #EDABE3, #CF4FE8, #BEECF4, #22AEC3, #9F92C8

---

[1] 22 hours
[2] [American Public Transportation Association]
[3] [U.S. Census Bureau]
[4] As described above
[5] We have checked this by inserting a picture of the visualizations into a color-blind tool online, and the colors were distinguishable from each other without much effort.

All of the hover text that appears when hovering over a line on a line chart or a bar on a bar chart has custom styling to ensure readability. Additionally, the text that appears is also customized for each graph to make sure that useful information is accessible.

We have added an annotation to the graphs that deal with yearly data, which highlights the beginning of the covid-19 pandemic.

4.2 Graphs

The dashboard consists of 4 pages, an introduction page, and then a page for each question posed in section 3. We have used 3 graphs to answer the first question, 2 graphs for the second and 3 for the last question. We have 4 different types of visualizations: Doughnut charts, bar charts, line charts and map charts.

To visualize the distribution of the different event types we have chosen to use a doughnut chart. The goal of this chart is also to introduce the reader to the dataset. The different event types are: Assault, Collision, Derailment, Fire, Other and Security.

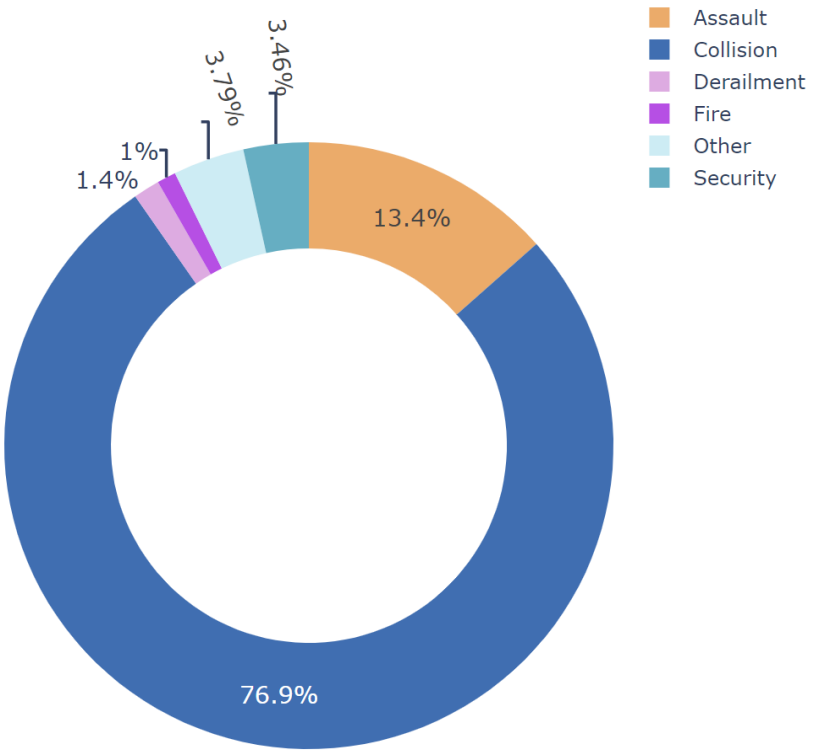

Figure 1: Doughnut chart

To show the development in the amount of occurrences of each event type and the development from 2014 to 2022, we use a line chart. This chart includes the annotation for the covid 19 pandemic, and the reader can interact with the chart, by zooming, and removing event types.

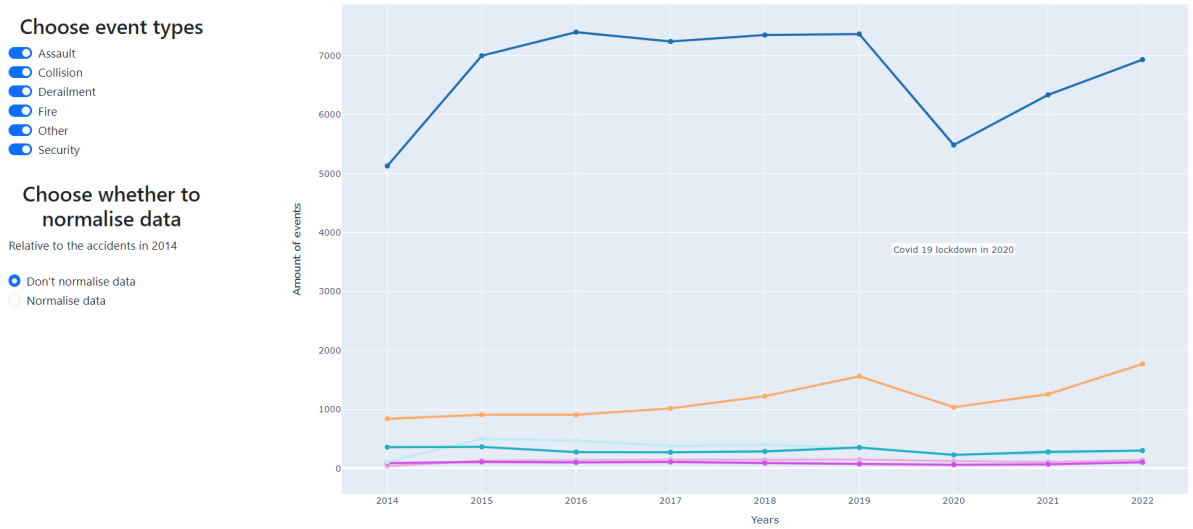

Figure 2: Line chart showing the development of different event types in the period 2014 to 2022. Legend present, but not shown.

Since the collision category dominates the chart, the graph can be normalized so every event type is indexed to the first year. This makes it possible to see the relative change in frequency over the years.

To reveal more insight into how the change of events differs for the different states in the United States, we have another line chart, which shows one event type at a time. The reader can choose what states to show.

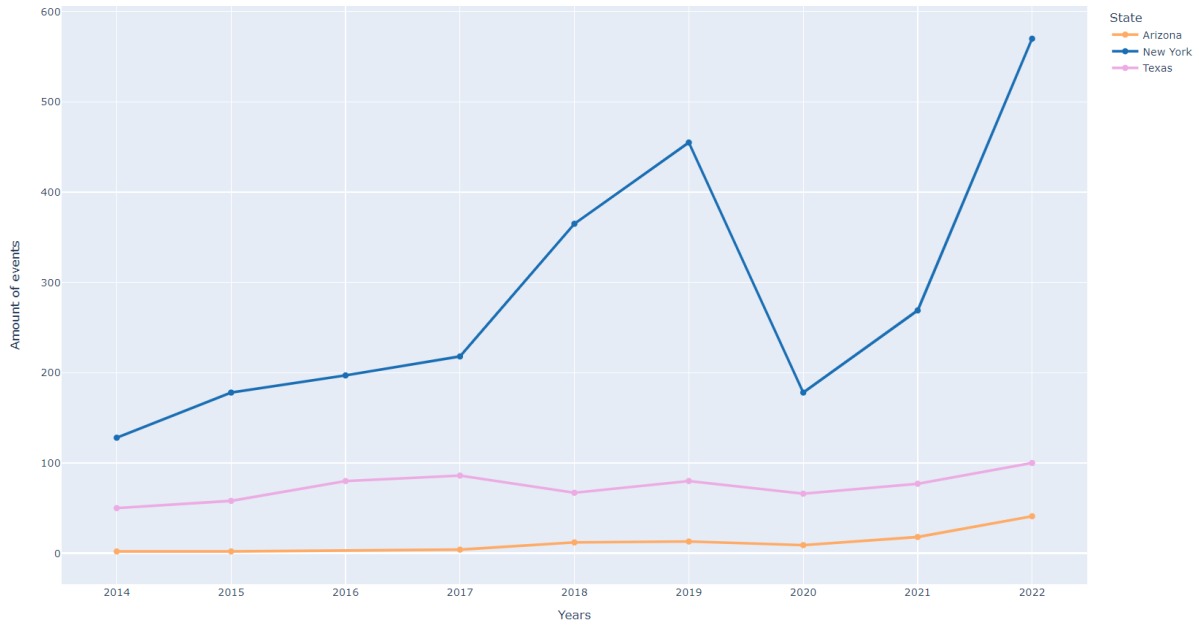

Figure 3: Line chart visualizing the development of the different event types over the years 2014-2022 for different states.

Furthermore, the graph allows for normalizing the data like the previous graph does. In addition, the user can select and deselect different states. The default is to only visualize Arizona, whereafter the user can add or remove states. The selection options can be seen in figure 4.

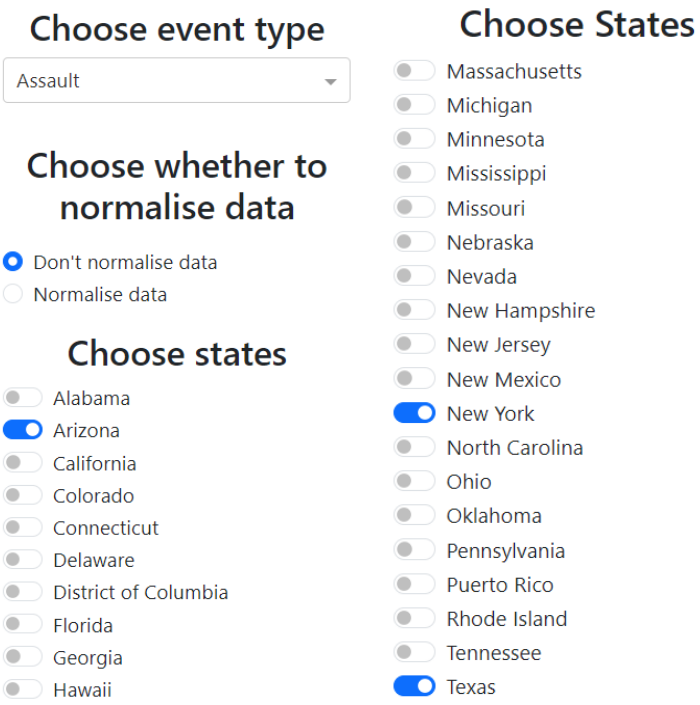

Figure 4: Interaction elements for figure 3

For question 2 we use a vertical bar chart that shows the time of day that incidents happen.

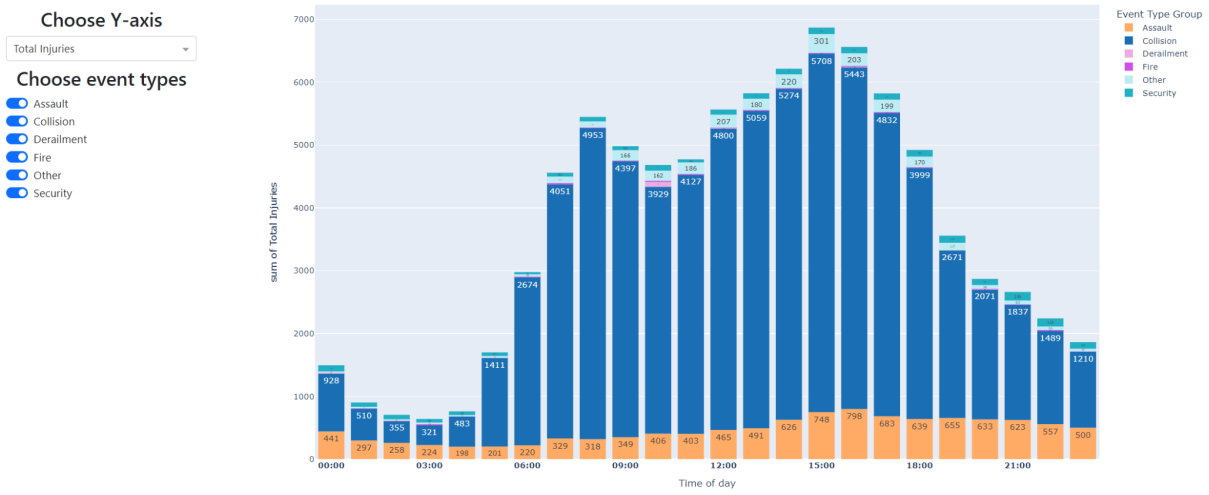

Figure 5: Vertical bar chart showing the number of events for each hour of the day.

This graph uses a stacked bar chart because we wish to convey the whole of the data. We want to show how many accidents happen at each time of day. By using the interaction elements at the left, the reader can investigate the change of each event type if they wish to do so.

To investigate how this distribution changes over the course of a year we create another stacked bar chart that only shows one month at a time. This chart can be animated which will play through the months starting with january.

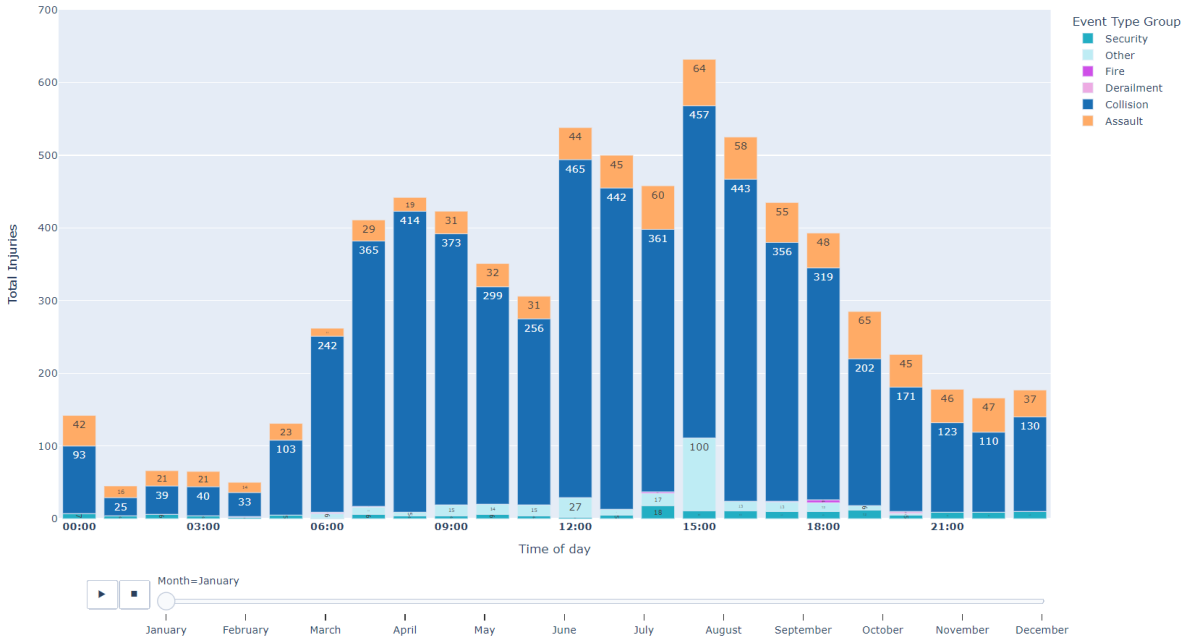

Figure 6: Vertical bar chart showing the change in distribution over the months.

Selecting and deselecting event types and choosing between total injuries, total fatalities and amount of accidents are offered on this graph as well. This chart sums up all of the years, into each month.

To show the locality of the safety events, and their geographic distribution we create a hexbin map. The map succeeds in this by laying hexagons over the map, wherein the number of accidents determines the color of the hexagon. The map can be seen in the following figure. In the major cities the number of accidents in a hexagon far exceeds that of the lesser populated regions. For this reason we have chosen to make a custom color scale, that is explained on the right of the chart, in two different versions.

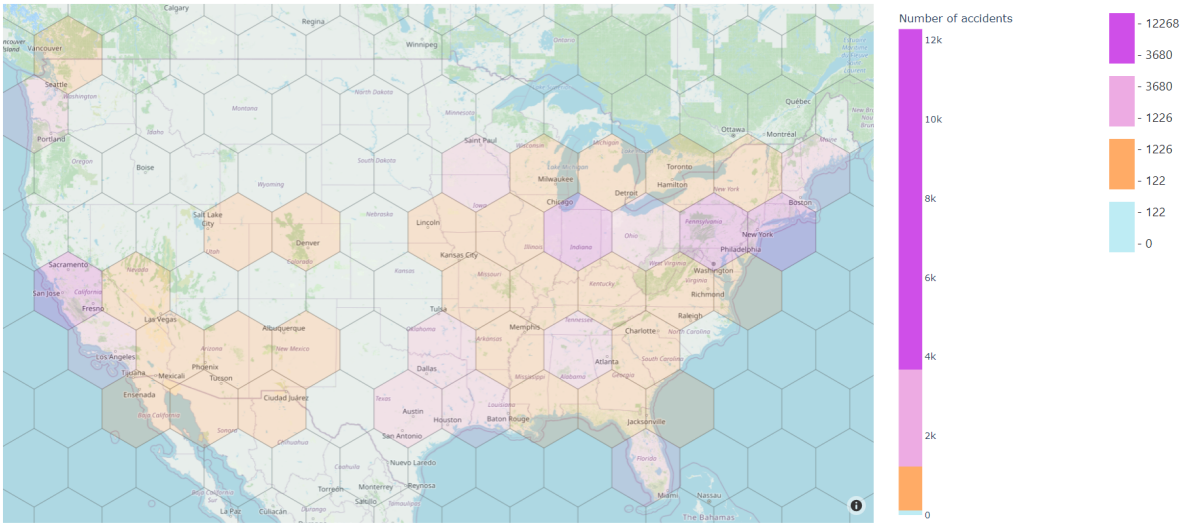

Figure 7: Hexbin map showing the places with the most accidents.

Hexagons in the size, used in figure 7, paints a very broad picture with little detail. Therefore, the users can adjust the amount of hexagons, to show more detail.

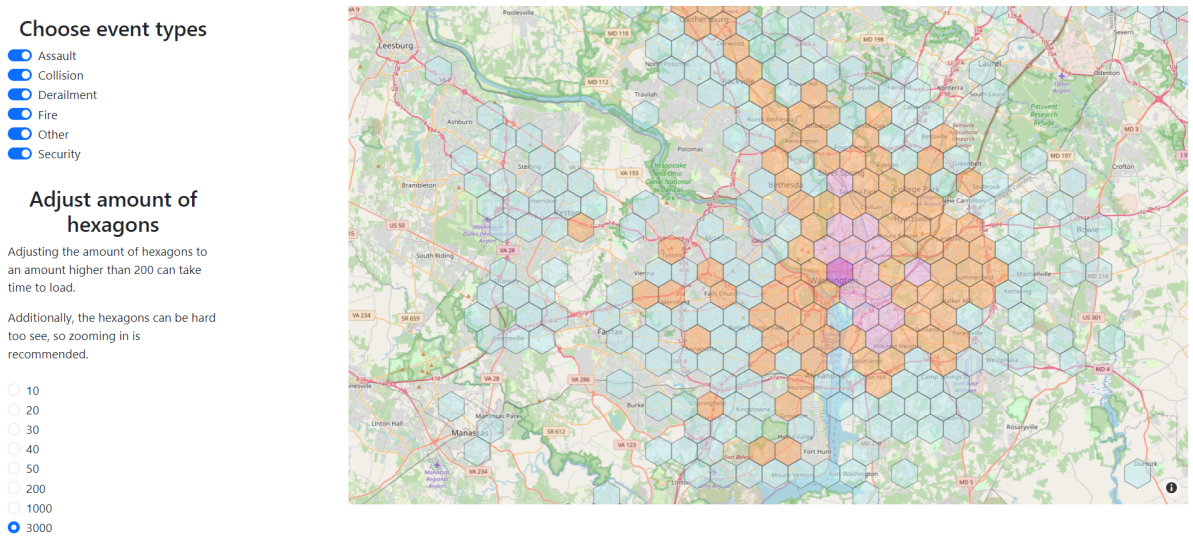

Figure 8: Hexbin map's interaction elements. The map is zoomed in on Washington.

The map allows the readers to adjust the amount of hexagons between 10 and 10000. Going beyond this takes too much time to load all the hexagons on the map.

We use a Hexbin Mapbox, because it shows the summation of data points. Furthermore, it is easier to see the areas with the most accidents, and it is easier to see the areas with the least accidents. The weakness of this plot is that it is not normalized to the amount of citizens in the area, which is why we see major hotspots in the largest cities. This problem is helped by the graph that comes after the next.

To show which states are dominated by different event types, we use a bar chart. This shows the 10 states that have the largest relative percentage of accidents of this type.

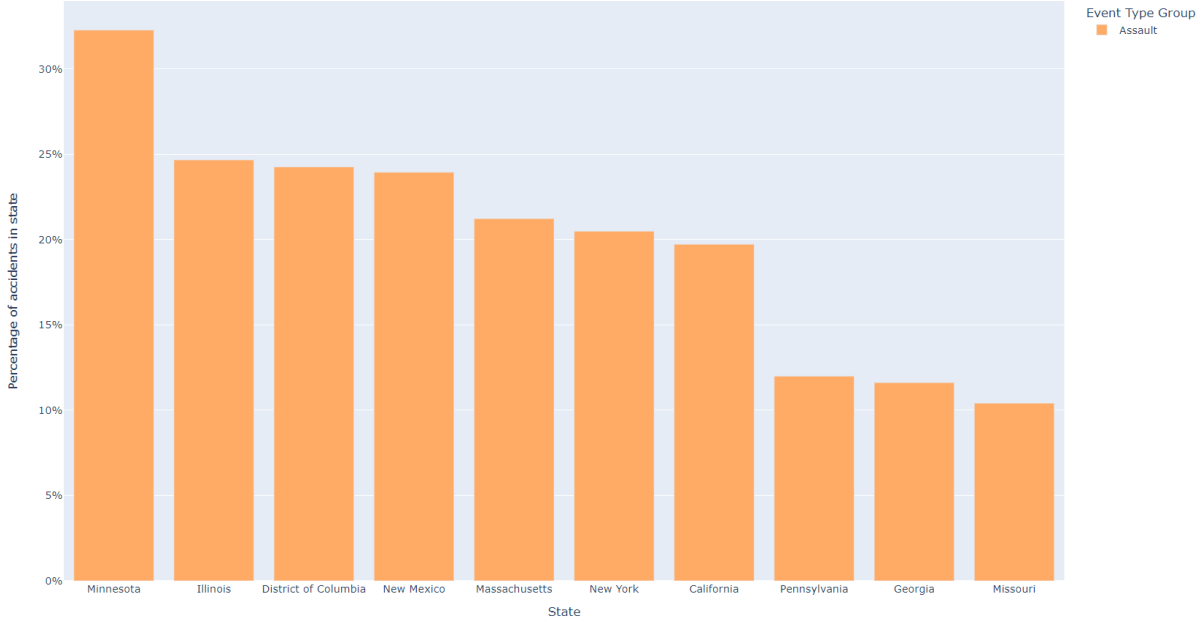

Figure 9: Bar chart over top 10 states, with the largest relative share of major safety incidents happening.

In this chart we use the same color for all states, because throughout the dashboard, these colors are used to indicate event types. As such the color will change when the reader selects a different event type.

To compare the number of accidents in a normalized fashion between the states, we calculate the number of events happening per million citizens in each state. We show this information in a stacked bar chart, because we want to compare the total events between the states.

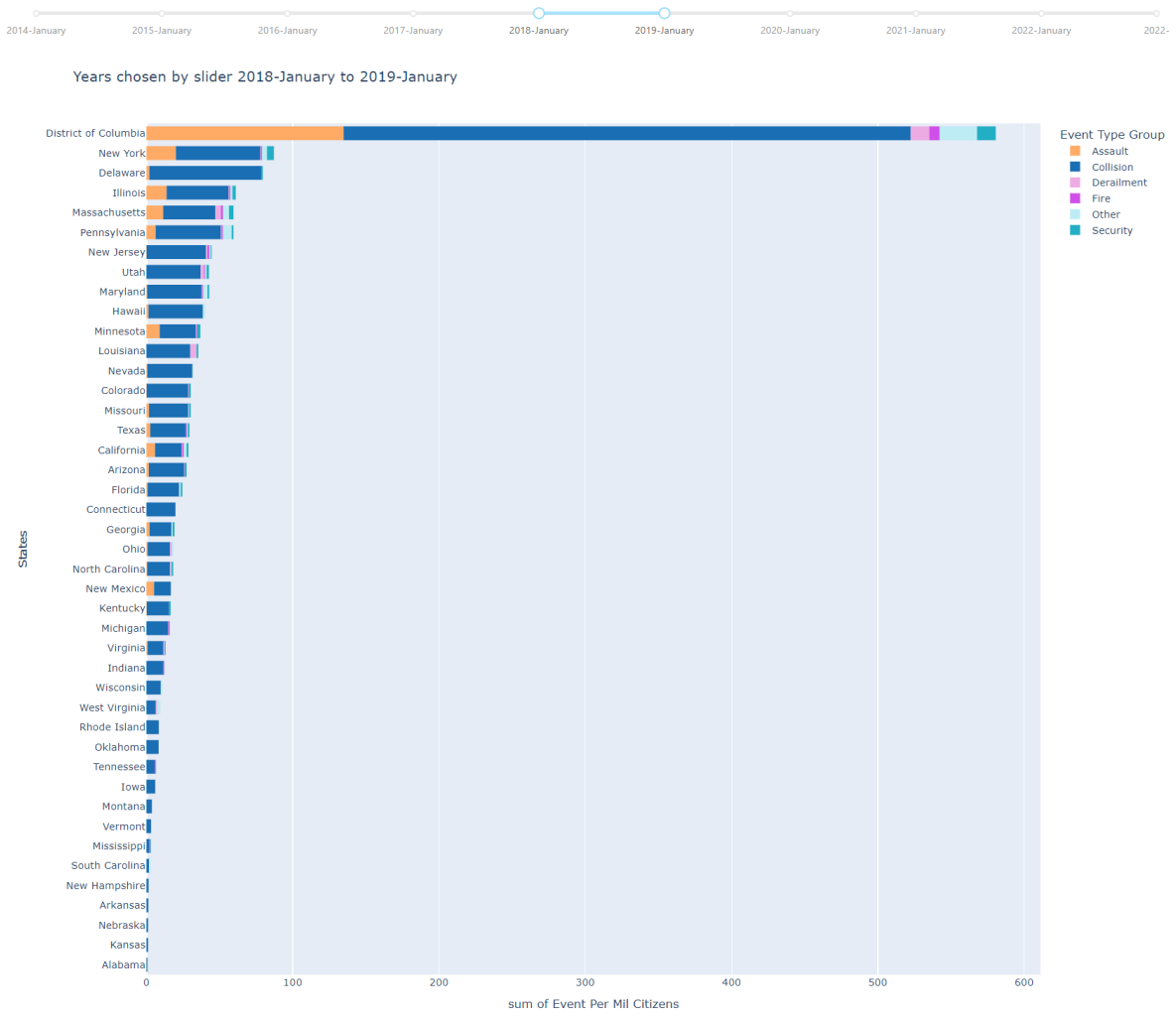

Figure 10: Horizontal bar graph visualizing the events divided by million citizens per state.

The graph can be adjusted to show a specific month or a range of months between January 2014 and December 2022. This graph shows the District of Columbia clearly at the top. This is a side effect of the fact that DC is a tiny state compared to the others, and it is very densely populated, which skews the data when adjusting for per-capita. To get a more accurate representation, this state should possibly be included in the data from one of the surrounding states.

Common for all of the graphs are that detailed information is shown when hovering over points on the line charts, bar charts or on a segment of the pie chart. This data aims at improving readability and understandability of the graphs and what the graphs are showing. These efforts further improve accessibility of the visualizations.

5. Story/Results

5.1 The Story

The visualizations tell the story of how major events in the public traffic have developed through the years from 2014 to 2022. By far the most common type of event is a collision. Examining the data reveals that for some states all types of collisions are reported, even minor fender-benders. The data shows a slow increase in overall events from 2014 to 2019, a drop in 2020, which was the year Covid-19 was introduced and an increase in 2021 and 2022 after the pandemic.

The development of major safety events differs for each state. Some states follow the overall development, while others are seeing major increases or decreases in certain types of events. Arizona, Florida and New York have seen an immense growth in assaults during 2021 and 2022, while Nevada and Colorado have seen a decline.

Looking at the time of the day where accidents happened. The data reveals that most public traffic events happen during business hours as one would expect. However, the total fatalities does not seem to follow the same curve. This is interesting information, and it would be relevant to explore this further. What could be the reason that the number of fatalities remains constant throughout the day? Unfortunately the data at hand does not provide enough insight for us to determine this. However, if the data is to be believed, the likelihood of an accident happening rises during the day, but these accidents are not as deadly as those happening at night.
These trends seem to stay relatively constant throughout the year, and we have not been able to determine a clear change in the times that accidents happen throughout the year. This Is somewhat contrary to our initial hypothesis, that accidents would move with the light, but perhaps due to modern society, the working hours do not change throughout the year.

For the different states there is a large difference in how many major safety events occur. Texas, California and New York are seeing the most events. However, events per million citizens shows that citizens in Washington D.C. and New York are more likely to be involved in a public transport major safety event

Washington D.C. has way more major safety events per million citizens compared to all the other states. Factors for this could be; that a large portion of people are living outside of the state, but working inside the state, that there could be reporting bias meaning safety officers in Washington D.C. report events more often, or that the data could be skewed. Our method of normalization might also be a large factor for this. Washington D.C. is a tiny state, with a high population density. Every square inch of the state is used, no other state is like it, which most likely skews the figures.

5.2 What surprised us?

Generally, the result is very close to expected besides the large outlier that was Washington D.C.'s amount of events per million citizens. It was expected that some high density areas would have an increased amount of events per citizen, yet Washington D.C. beats all the other states by a mile.

Furthermore, it was surprising to see that the amount of assaults have doubled between 2014 and 2022 while collisions only grew by around 35 percent. Many different factors could have had an effect on this, yet it is still a huge growth. The reason for this would be interesting to know.

It was also surprising to discover that the number of fatalities stays constant throughout the entire day.

5.3 What did we learn about the data?

From the question: "Do certain types of accidents occur more in certain types of environments", we learned that 36 percent of West Virginia's major safety events are of the type "Other". It seems unlikely that West Virginia has such a large share of incidents in this category, so maybe the people responsible for data entry have gotten used to filing events in the Other category if they seem slightly out of place. It definitely seems odd that 36 percent of West Virginia's accidents are in that category, when the second highest state in that category placed 8.84 percent.
It seems likely that not all of the states follow the same rules/guidelines when reporting major safety events, skewing the data. This can be seen from thousands of records missing latitude and longitude and some states having an absurd amount of specific event types or missing event types entirely as seen in a third of West Virginia's events being of type "Other". It is also clear that some states apparently do not use the system at all, since at least two states have no records at all.

5.4 How well did the visualization work, and how could it be further improved?

Overall, the visualizations work well. The color scheme applied to the dashboard as well with how modifiable the graphs are makes the dashboard interesting and allows the reader to explore the dataset.

Some graphs could still be further improved to improve the dashboard. The line chart on the page regarding the question: *"1. Is there a relation between time periods in the day and certain types of accidents?"* uses color to differentiate between different states selected. This could be improved by having all the lines gray and highlighting the line selected or hovered on by the user. This would improve the graph if more than six states are present on the line chart, since the color scheme only has seven colors.
We could also change the top-10 chart by creating a combination of pie chart and map chart, by overlaying the top 10 states, showing their share as a pie chart on top of the states location on the map. This would engage the reader more, and possibly make it faster to spot

the changing states. The idea for this chart is to only show the pies with the chosen event type filled out (i.e. filling the pie for minnesota with 32% Assault, and the rest gray if assault is selected).

The dashboard could be further improved for the color blind users by introducing more identifiers than color to the equation. One of these identifiers could be having different geometric objects for the types of events such that collisions have a square attached to them and assaults have a triangle attached. These objects could then be shown when hovering over lines on the line charts and bars on the bar charts.

6. Conclusion/Discussion

In conclusion, the dashboard succeeds in answering the following three questions:

1. Is there an increase or decrease in certain types of accidents in the last 9 years?
2. Is there a relation between time periods in the day and certain types of accidents?
3. Do certain types of accidents occur more often in certain environments?

The dashboard gives the readers insights into these questions through eight graphs; four barcharts, two line charts, one hexbin map and a doughnut chart. In addition, the reader is guided through the visualization by the accompanying text. A focus on interactivity and colorblind friendliness was applied resulting in an interactive, readable dashboard.

Regarding the growth and decline in certain types of accidents in the last 9 years, the visualizations reveal that most of the different event types have seen growth, while “Security” has seen a small decline. Derailment, Other and Assault have all more than doubled since 2014.

There seems to be some relation between time of day and the different types of major safety events. “Collisions” and “Other” events tend to happen during business hours, while “Assaults” and “Security” events tend to happen later in the evening. Despite this correlation, it seems that fatalities happen at the same rate throughout the entire 24 hours of a day.

Overall most major safety events occur in states with larger cities. On top of that, these states tend to have a higher percentage of major safety events being of type “Assault”. West Virginia has way more events of type “Other” than all the other states, whereof the legitimacy of can be questioned.

During development of the dashboard, many challenges were faced regarding the dataset. The dataset was cleaned by removing numerous empty columns, and precomputing some new columns specific for our use case.

7. Participation

Student	PoC
Tobias Vittrup Bak, tobib21@student.sdu.dk	100
Thor Malmby Jørgin, tjoer21@student.sdu.dk	100
Kevin Torp, ketor21@student.sdu.dk	94.279
Philip Schwartz, pschw21@student.sdu.dk	30

References:

U.S. Census Bureau. 2023. Population Estimates Program Datasets. Retrieved 22/12/2022, from https://www2.census.gov/programs-surveys/popest/datasets/

American Public Transportation Association. 2020. National Transit Database Tables. File used: “2020 Agency Information”. Retrieved 22/12/2022, from https://www.apta.com/research-technical-resources/transit-statistics/ntd-data-tables/